Title
Bayesian selection of nucleotide substitution models and their site assignments.

Journal
Molecular biology and evolution, 30(3)

ISSN
0737-4038

Authors
Wu, Chieh-Hsi
Suchard, Marc A
Drummond, Alexei J

Publication Date
2013-03-01

DOI
10.1093/molbev/mss258

Peer reviewed

Bayesian Selection of Nucleotide Substitution Models and Their Site Assignments

Chieh-Hsi Wu,[1,2] Marc A. Suchard,[3,4,5] and Alexei J. Drummond*[,1,2]

[1]Department of Computer Science, University of Auckland, Auckland, New Zealand

[2]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

[3]Department of Biomathematics, David Geffen School of Medicine at UCLA

[4]Department of Human Genetics, David Geffen School of Medicine at UCLA

[5]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles

***Corresponding author:** E-mail: alexei@cs.auckland.ac.nz.

Associate editor: Jeffrey Thorne

Abstract

Probabilistic inference of a phylogenetic tree from molecular sequence data is predicated on a substitution model describing the relative rates of change between character states along the tree for each site in the multiple sequence alignment. Commonly, one assumes that the substitution model is homogeneous across sites within large partitions of the alignment, assigns these partitions a priori, and then fixes their underlying substitution model to the best-fitting model from a hierarchy of named models. Here, we introduce an automatic model selection and model averaging approach within a Bayesian framework that simultaneously estimates the number of partitions, the assignment of sites to partitions, the substitution model for each partition, and the uncertainty in these selections. This new approach is implemented as an add-on to the BEAST 2 software platform. We find that this approach dramatically improves the fit of the nucleotide substitution model compared with existing approaches, and we show, using a number of example data sets, that as many as nine partitions are required to explain the heterogeneity in nucleotide substitution process across sites in a single gene analysis. In some instances, this improved modeling of the substitution process can have a measurable effect on downstream inference, including the estimated phylogeny, relative divergence times, and effective population size histories.

Key words: across-site rate variation, Dirichlet process mixture model, Bayesian model selection.

Introduction

Phylogenetic analysis in a probabilistic framework requires the adoption of a substitution model. However, much uncertainty lingers about modeling this process. For example, which substitution model is most suitable for the analysis given the data set and how does the substitution process vary across sites? It is well established that substitution rates exhibit variation across sites (Yang 1996) and omitting across-site rate variation can result in inaccurate estimation of the phylogeny (Huelsenbeck and Hillis 1993) and underestimation of branch lengths if substitutions occur repeatedly at sites undergoing rapid evolution (Sullivan and Joyce 2005). Incorporating across-site variation in the underlying substitution model parameters themselves may improve the accuracy of phylogenetic parameter estimates (Huelsenbeck and Nielsen 1999). These parameters include the relative exchange rates between nucleotide character states and their stationary distribution. We use the term "substitution pattern" to refer to a particular set of restrictions among the values of these parameters. Differing restrictions lead to different named substitution models. How to select an appropriate substitution pattern and rate for all sites in an alignment remains a daunting task (Suchard et al. 2001).

One approach to relax the assumption of rate constancy across sites treats the overall rate multiplier at each site as a random variable distributed according to an underlying distribution shared across sites (Golding 1983; Jin and Nei 1990; Yang 1993). The most popular distribution is a discretized version of the Gamma distribution with a single shape parameter α (Yang 1994), but other distributions have also been explored (Olsen 1987; Waddell and Steel 1997). Another common modeling assumption is that some proportion of the sites are invariant (Hasegawa et al. 1985; Churchill et al. 1992; Waddell and Penny 1996). It has become common to use both a mixing distribution and a zero-inflation via this proportion of invariant sites to model the rate variation across sites (Gu et al. 1995; Waddell and Steel 1997). An alternative approach places the sites into categories and independently estimates the rate multiplier of each category. The most extreme partition scheme estimates a multiplier independently for each site (Swofford et al. 1996; Nielsen 1997), but this tends to vastly overfit the data, leading to undesirable statistical properties (Felsenstein 2004). The most common a priori partition scheme for protein coding genes is by codon position, with the estimated multiplier at the third codon position usually higher than those in the first

Article

and second codon positions due to redundancy in the genetic code. Other biologically reasonable partition schemes may also be appropriate (e.g., loop versus stem in RNA coding genes, or exposed versus buried region for amino acid sequences where 3D structure is known), but they are not easy to determine. A Bayesian nonparametric method, which employs a Dirichlet process mixture (DPM) model, enables the joint estimation of the number of rate categories and the site-to-category assignment (Huelsenbeck and Suchard 2007).

The across-site variation of relative exchange rates and the stationary distribution are, however, less often accounted for in most phylogenetic analyses. For nucleotides, Huelsenbeck and Nielsen (1999) have modeled variation in the transition/transversion rate ratio through a discretized gamma distribution (Huelsenbeck and Nielsen 1999). For amino acids, several partition schemes have been explored for amino acid substitution patterns across sites. The partition scheme by Bruno (1996) allows each site to have its own amino acid substitution pattern. Similar to the site independence in overall rate multiplier counterpart, such a scheme is likely to be subject to overfitting. Others have proposed partitions which first predefine 8–10 categories (Goldman et al. 1998; Koshi et al. 1999; Li and Goldman 1999; Dimmic et al. 2000; Soyer et al. 2002), where the categorization of some is based on protein features such as the secondary structure and solvent accessibility of the protein (Goldman et al. 1998; Li and Goldman 1999). Quang et al. (2008) have developed a method that estimates a mixture of a predetermined number of amino acid patterns from alignment databases via an expectation–maximization algorithm. As with partitioning sites for rate multipliers, it is often not obvious how many categories of amino acid patterns are required a priori. The CAT model (Lartillot and Philippe 2004) avoids this problem by using a DPM model. The DPM model has also been applied to model the variation in rate of nonsynonymous substitution across sites to detect positive selection (Huelsenbeck et al. 2006).

To judge the uncertainty of nucleotide substitution model selection, it has become almost standard procedure in recent years to first assign a named model to each predefined partition by ModelTest (Posada and Crandall 1998) before performing a more complex analysis in a different framework. In a Bayesian framework, an alternative to this two-step scheme is to use techniques that perform model selection and phylogenetic parameter estimation simultaneously. As single partition examples, Suchard et al. (2001) and Huelsenbeck et al. (2004, implemented in Ronquist et al. 2012, MrBayes 3.2), exploit reversible jump Markov chain Monte Carlo (Green 1995) to simultaneously select substitution models. Wu and Drummond (2011) have used a product space formulation of transdimensional MCMC (Godsill 2001) for selection of microsatellite mutation models. Lemey et al. (2009) have modeled the migration history of RNA viruses using continuous time Markov chains (CTMC) and applied "spike-and-slab" priors that provide nonzero probability mass on parameter restrictions for selection (Kuo and Mallick 1998) to infer the transmission route. Huelsenbeck et al. (2008) considered a general-time reversible parameterization of amino acid substitutions and all its submodels (i.e., some relative rate entries share the same value) as partitionings under a DPM model for selection.

In this article, we present a spike-and-slab-based mixture model for nucleotide alignment data that accounts for across-site heterogeneity of substitution pattern and rate multiplier simultaneously. It enables Bayesian selection over a set of standard nucleotide substitution models for each substitution model category. The assignment of sites to categories has a prior probability defined by the Dirichlet process (Ferguson 1973; Antoniak 1974). Under the Dirichlet process, both the category assignment and the number of categories are random variables. This nonparametric process is therefore a popular approach for problems where the data are thought to come from a mixture of an unknown number of probability distributions. We present two variants: the substitution Dirichlet mixture model 1 (SDPM1) specifies that the substitution pattern and rate multiplier share a common partitioning scheme and the substitution Dirichlet mixture model 2 (SDPM2) provides independent Dirichlet process priors for the pattern and rate multipliers. A recently proposed method by Lanfear et al. (2012, PartitionFinder) uses a greedy heuristic algorithm to find the partition that maximizes the likelihood for a given alignment. One main difference to our approach is that this method does not quantify the uncertainty associated with alignment partitioning. Also, our method produces phylogenies and population histories integrated over the space of alignment partitions and substitution model assignments.

Materials and Methods

The Model

To develop our SDPM1 and SDPM2 models, we start with a nucleotide sequence alignment $\boldsymbol{D}$ that consists of n taxa and s sites. The nucleotide pattern at site i is denoted as $\boldsymbol{D_i}$. For two sites i and j where $i \neq j$, they refer to different columns of the alignment and are treated as distinct entities whether or not their patterns are identical. $\boldsymbol{D}$ is assumed to be generated by an underlying CTMC, along a rooted bifurcating tree τ, representing an unknown phylogeny. The substitution process is determined by the rate multipliers $\boldsymbol{r} = \{r_1, \ldots, r_s\}$ and the substitution model parameters $\boldsymbol{\Phi} = \{\boldsymbol{\phi_1}, \ldots, \boldsymbol{\phi_s}\}$ across sites. Each $\boldsymbol{\phi_i}$ includes all the parameters that make up the infinitesimal rate matrix of CTMC at site i. In a Bayesian phylogenetic analysis, we seek the joint posterior distribution

$$f(\tau, \boldsymbol{\Phi}, \boldsymbol{r}|\boldsymbol{D}) \propto f(\boldsymbol{D}|\tau, \boldsymbol{\Phi}, \boldsymbol{r}) f(\tau) f(\boldsymbol{\Phi}, \boldsymbol{r}), \quad (1)$$

where the term $f(\tau)$ is the prior density on the tree and $f(\boldsymbol{\Phi}, \boldsymbol{r})$ is the joint prior density over the evolutionary model parameters. Here, we assume prior independence between the tree and evolutionary model parameters. If we apply a coalescent prior to the tree, then $f(\tau)$ is replaced by $f(\tau|\boldsymbol{\Theta}) f(\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ contains the demographic parameters of the coalescent and has hyperprior density $f(\boldsymbol{\Theta})$. The term $f(\boldsymbol{D}|\tau, \boldsymbol{\Phi}, \boldsymbol{r})$ is the likelihood given all model parameters. The likelihood at site i, $f(\boldsymbol{D_i}|\tau, \boldsymbol{\phi_i}, r_i)$, is calculated by

Felsenstein's pruning algorithm (Felsenstein 1981), and the full likelihood is the product of the likelihood over all sites:

$$f(\boldsymbol{D}|\tau,\boldsymbol{\Phi},\boldsymbol{r}) = \prod_{i=1}^{s} f(\boldsymbol{D_i}|\tau,\boldsymbol{\phi_i},r_i). \quad (2)$$

Heterogeneity of Evolutionary Parameters across Sites

If the evolutionary process is homogeneous across sites then $r_1 = r_2 = \ldots = r_s$ and $\boldsymbol{\phi_1} = \boldsymbol{\phi_2} = \ldots = \boldsymbol{\phi_s}$. To relax this assumption, we estimate an unknown partitioning of the evolutionary model parameters across sites using DPM models.

Consider the SDPM1 model wherein the substitution model parameters and rates share the same partitioning. Let K be an unknown parameter denoting the number of categories of evolutionary model parameters. The substitution model parameters and rate at each site are assigned to one of the K categories. Each category has its own unique set of values of evolutionary model parameters. Let $\boldsymbol{\Phi}^*$ be the union of unique substitution model parameters over all categories, whereas $\boldsymbol{r}^*$ is the union of unique rate multipliers values across all categories. The term σ_i denotes the category to which site i has been assigned, where $\sigma_i \in \{1, \ldots, K\}$, therefore $\boldsymbol{\phi_i} = \boldsymbol{\Phi}^*_{\sigma_i}$ and $r_i = r^*_{\sigma_i}$. We can rewrite equation (1) in terms of $\boldsymbol{\Phi}^*$, $\boldsymbol{r}^*$, and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_s)$, such that

$$\begin{aligned} f(\tau,\boldsymbol{\Phi},\boldsymbol{r}|\boldsymbol{D}) &= f(\tau,\boldsymbol{\Phi}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}|\boldsymbol{D}) \\ &\propto f(\boldsymbol{D}|\tau,\boldsymbol{\Phi}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}) f(\tau) f(\boldsymbol{\Phi}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}). \end{aligned} \quad (3)$$

Under the Dirichlet process,

$$f(\boldsymbol{\Phi}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}) = \frac{\chi^K \prod_{k=1}^{K} (\psi_k - 1)!}{\prod_{i=1}^{s} (\chi + i - 1)} \prod_{k=1}^{K} G_0^{\Phi}(\boldsymbol{\Phi}^*_{\boldsymbol{k}}) G_0^{r}(r^*_k), \quad (4)$$

where ψ_k is the number of sites assigned to category k, distributions G_0^{Φ} and G_0^{r} are the base distributions of substitution model parameters and rate multipliers, respectively, and $\chi \in (0, \infty)$ is the "concentration parameter" of the Dirichlet process. Notice that permutation of the assignment vector $\boldsymbol{\sigma}$ does not affect the distribution in equation (4). Parameter χ controls the marginal distribution on the number of categories a priori:

$$f(K|\chi, s) = \frac{S_1(s,K)\chi^K}{\prod_{i=1}^{s} (\chi + i - 1)}, \quad (5)$$

where $S_1(s, K)$ is the absolute value of the Stirling number of the first kind given parameter values s (number of sites) and K. According to equation (5), the Dirichlet process tends to produce more categories with increasing χ.

If the substitution model parameters and rates across sites are modeled by independent Dirichlet processes as in the SDPM2 model, then the full posterior can be written as follows:

$$\begin{aligned} &f(\tau,\boldsymbol{\Phi}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}^{\boldsymbol{\Phi}},\boldsymbol{\sigma}^{\boldsymbol{r}}|\boldsymbol{D}) \propto \\ &f(\boldsymbol{D}|\tau,\boldsymbol{\Phi}^*,\boldsymbol{r}^*,\boldsymbol{\sigma}^{\boldsymbol{\Phi}},\boldsymbol{\sigma}^{\boldsymbol{r}}) f(\tau) f(\boldsymbol{\Phi}^*,\boldsymbol{\sigma}^{\boldsymbol{\Phi}}) f(\boldsymbol{r}^*,\boldsymbol{\sigma}^{\boldsymbol{r}}), \end{aligned} \quad (6)$$

where $\boldsymbol{\sigma}^{\boldsymbol{\Phi}}$ and $\boldsymbol{\sigma}^{\boldsymbol{r}}$ are the respective assignment vectors for the substitution model parameters and rates. The prior distribution of substitution model parameters across sites is as follows:

$$\begin{aligned} f(\boldsymbol{\Phi}) &= f(\boldsymbol{\Phi}^*,\boldsymbol{\sigma}^{\boldsymbol{\Phi}}) \\ &= \frac{\chi(\Phi)^{K^{\Phi}} \prod_{k=1}^{K^{\Phi}} (\psi_k^{\phi} - 1)!}{\prod_{i=1}^{s} (\chi(\Phi) + i - 1)} \prod_{k=1}^{K^{\Phi}} f_0^{\phi}(\boldsymbol{\Phi}^*_{\boldsymbol{k}}), \end{aligned} \quad (7)$$

where ψ_k^{ϕ} is the number of sites assigned to category k of the K^{Φ} substitution model categories, and $\chi(\Phi)$ is the concentration parameter of the Dirichlet process prior on substitution model partition. The prior distribution of rates across sites follows similarly. We let ψ_k^{r} denote the number of sites in category k of the K^{r} rate categories and $\chi(r)$ denote the concentration parameter of the Dirichlet process prior on rate partition.

Posterior Inference of Partitioning

We employ a Gibbs sampling procedure (Neal 2000, algorithm 8) for updating the assignment vector $\boldsymbol{\sigma}$ in the SDPM1 model. Site i, which is in category k ($\sigma_i = k$), is picked randomly and removed from the rest of the sites. If there are currently K classes, let K^{-i} denote the number of categories after the removal of site i. If site i is an singleton, we create κ auxiliary sets of substitution model parameters and rates by setting $K^{-i} + 1$ to k and draw new parameter values from the base distribution for each of the categories in $\{K^{-i} + 2, \ldots, K^{-i} + \kappa\}$. If site i is not a singleton, a new set of evolutionary model parameters are drawn for each of the κ auxiliary categories. The Gibbs sampler proposes a new category, σ_i' with probability

$$f(\sigma_i' = k') = \begin{cases} h\psi_{k'}^{-i}\ell(\boldsymbol{D_i}) & \text{if } 1 \le k' \le K^{-i} \\ h\frac{\chi}{\kappa}\ell(\boldsymbol{D_i}) & \text{if } K^{-i} < k' \le K^{-i} + \kappa, \end{cases} \quad (8)$$

where $\ell(\boldsymbol{D_i}) = f(\boldsymbol{D_i}|\tau,\boldsymbol{\phi}_{k'},\boldsymbol{r})$ and h is the normalizing constant. Categories are discarded if they are not associated with any site after the update. For the analyses in this study, we use $\kappa = 5$. The same procedure is used to update $\boldsymbol{\sigma}^{\boldsymbol{\Phi}}$ and $\boldsymbol{\sigma}^{\boldsymbol{r}}$ in SDPM2.

The Gibbs sampling procedure described above updates the assignment vector site-by-site and therefore lacks efficiency when the number of sites is large because site assignments are highly correlated. To overcome this issue, we also employ a Metropolis–Hastings (Metropolis et al. 1953; Hastings 1970) sampling algorithm that makes updates of assignment at multiple sites in one step by splitting and merging existing categories (Dahl 2005). Using $\boldsymbol{\sigma}$ as an example, a sequentially allocated-split-merge sampling has the following steps. We randomly choose a pair of sites i and j, where $i \neq j$. If i and j are in the same category k, then k will be split. After removing sites i and j from k, we let $S(k^{-\{i,j\}})$ denote the set of sites associated with k without i and j. We can then construct two new categories, $k^{(i)}$ containing site i and $k^{(j)}$ containing site j. We draw one site, u, at a time

without replacement from $S(k^{-\{i,j\}})$ and assign it to $k^{(i)}$ with the probability

$$\Pr(\sigma_u' = k^{(i)}) = \frac{\psi_{k^{(i)}} f(D_u|\boldsymbol{\tau}, \phi_{k^{(i)}}, \boldsymbol{r})}{\psi_{k^{(i)}} f(D_u|\boldsymbol{\tau}, \phi_{k^{(i)}}) + \psi_{k^{(j)}} f(D_u|\boldsymbol{\tau}, \phi_{k^{(j)}})}. \quad (9)$$

The model parameters $\phi_{k^{(i)}}$ and $r_{k^{(i)}}$ are updated by drawing values from their respective base distributions. After each allocation of u, either $\psi_{k^{(i)}}$ or $\psi_{k^{(j)}}$ increments by 1. The proposal density of splitting a category is the product of equation (9) after each draw from $S(k^{-\{i,j\}})$ multiplied by $G_0^{\Phi}(\phi_{k^{(i)}})G_0^{r}(r_{k^{(i)}})$. The proposal probability of the reversal step is 1.0, as there is only one assignment option to merge two categories.

If sites i and j are in different categories, k^i and k^j, respectively, then they are merged into one category, say k^m. The parameter values associated with this category are set to $\boldsymbol{\phi}_{\boldsymbol{k}}^{\boldsymbol{j}}$. The proposal probability of a merge step is 1.0. The reverse proposal probability is $G_0(\phi_{k^{(i)}})$ multiplied by the product of the probabilities in equation (9) for an assignment choice required to obtain the split allocation to k^i and k^j from the merged category k^m.

Bayesian Model Selection

We use a spike-and-slab prior specification (Kuo and Mallick 1998) to facilitate Bayesian selection among named nucleotide substitution models for each category. Under this approach, we augment ϕ_r to include a set of binary indicator variables, whose realized 0, 1 values allow us to move between substitution model parameter restrictions that correspond to common nucleotide models. Specifically, the infinitesimal rate matrix of category k is $Q_k = \Lambda_k \Pi_k$, where Λ_k is a symmetric matrix with upper-triangular entries

$$\Lambda_k = \begin{pmatrix} - & \phi_{k,\mathrm{AC}} & \phi_{k,\mathrm{AG}} & \phi_{k,\mathrm{AT}} \\ & - & \phi_{k,\mathrm{CG}} & \phi_{k,\mathrm{CT}} \\ & & - & \phi_{k,\mathrm{GT}} \\ & & & - \end{pmatrix} \quad (10)$$

and matrix Π_k is diagonal with entries $(\pi_{k,\mathrm{A}}, \pi_{k,\mathrm{C}}, \pi_{k,\mathrm{G}}, \pi_{k,\mathrm{T}})$. Using the binary indicators $\boldsymbol{\delta}_k = (\delta_{k,TN}, \delta_{k,\kappa}, \delta_{k,TV}, \delta_{k,f_k})$, we further parameterize

$$\begin{aligned} \log \phi_{k,\mathrm{AG}} &= 0 \\ \log \phi_{k,\mathrm{CT}} &= \delta_{k,\mathrm{TN}} \rho_{k,\mathrm{TN}} \\ \log \phi_{k,\mathrm{AC}} &= -\delta_{k,\kappa}\rho_{k,\kappa} + \delta_{k,\mathrm{TV}}\rho_{k,\mathrm{AC}} \\ \log \phi_{k,\mathrm{AT}} &= -\delta_{k,\kappa}\rho_{k,\kappa} + \delta_{k,\mathrm{TV}}\rho_{k,\mathrm{AT}} \\ \log \phi_{k,\mathrm{GC}} &= -\delta_{k,\kappa}\rho_{k,\kappa} + \delta_{k,\mathrm{TV}}\rho_{k,\mathrm{GC}} \\ \log \phi_{k,\mathrm{GT}} &= -\delta_{k,\kappa}\rho_{k,\kappa}, \text{ and} \\ \log \pi_{k,b} &= (1-\delta_{k,\mathrm{FQ}})\log(1/4) + \delta_{k,\mathrm{FQ}} \log f_{k,b}, \end{aligned} \quad (11)$$

for $b \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Each element of $\boldsymbol{\rho}_k = (\rho_{k,\mathrm{TN}}, \rho_{k,\kappa}, \rho_{k,\mathrm{AC}}, \rho_{k,\mathrm{AT}}, \rho_{k,\mathrm{GC}})$ takes a value in the range $(-\infty, \infty)$. The base frequencies $f_k = (f_{k,\mathrm{A}}, f_{k,\mathrm{C}}, f_{k,\mathrm{G}}, f_{k,\mathrm{T}})$ satisfy $0 \le f_{k,b} \le \sum_b f_{k,b} = 1$. When certain indicators in $\boldsymbol{\delta}_k$ achieve the value 0, specific effects fall out of the model. Using this approach, we are able to conveniently parameterize the Kimura (1980, K80), Felsenstein (1981, F81), Hasegawa et al. (1985, HKY85), Tamura and Nei (1993, TN93), and Tavaré (1986, general time reversible [GTR]) infinitesimal rate matrices. Table 1 presents the relationship between $\boldsymbol{\delta}_k$ and these named models. Also presented in the table 1 is an alternative parameterization of $\boldsymbol{\delta}_k$ into a single categorical variable ξ_k achieving five partially ordered values. ξ_k takes values K80, F81, HKY85, TN93, and GTR. Sampling ξ_k provides an opportunity to traverse through substitution model space without changing the total model dimension. Finally, the infinitesimal matrix $\boldsymbol{Q}_k$ is normalized, so that the total mutational outflow is 1.0; in other words we multiply $\boldsymbol{Q}_k$ by $c = -1/\sum_b \pi_b q_{bb}$.

Table 1. Indicator Values of a Given Substitution Model.

Indicator	Substitution Model (ξ)				
	K80	F81	HKY85	TN93	GTR
δ_{FQ}	0	1	1	1	1
δ_{κ}	1	0	1	1	1
δ_{TN}	0	0	0	1	1
δ_{TV}	0	0	0	0	1

Single-Locus Data

We applied our method to four single-locus data sets of gene coding sequences, three of which are collected from RNA viruses and one from mammalian species.

Ebola Virus

The Ebola virus (EBOV) data set was compiled by Wertheim and Kosakovsky Pond (2011). It consists of 32 glycoprotein sequences of 1,389 base pairs. The sampling times range from 1976 to 2005.

Hepatitis C Subtype 4

The hepatitis C subtype 4 (HCV-4) data set was data set B in a study on the population genetics and epidemiology history of HCV in Egypt (Pybus et al. 2003). It was originally from a comprehensive study on the diversity of HCV in Egypt (Ray et al. 2000). This data set contains 63 contemporaneous sequences of 411 base pairs from the E1 region.

Mammal

The mammal data set was obtained from the OrthoMam database (Ranwez et al. 2007). The data set contains sequences from 12 mammalian species: *Canis familiaris*, *Felis catus*, *Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus abelii*, *Macaca mulatta*, *Microcebus murinus*, *Otolemur garnettii*, *Mus musculus*, *Rattus norvegicus*, *Ochotona princeps*, and *Oryctolagus cuniculus*. The sequences have length of 468 base pairs and are from *FGF1* gene, which codes for heparin-binding growth factor 1.

Respiratory Syncytial Virus Subgroup A

This data set has 35 sequences and 629 sites from the *G* gene of the human respiratory syncytial virus subgroup A (RSVA) sampled from 1956 to 2002 (Zlateva et al. 2005).

Hepatitis C Virus Subtype 1b Full-Genome Data

We also analyzed a data set of HCV subtype 1-b genomes used in the study by Gray et al. (2011). It consists of 31 within-host sequences of 9,030 sites sampled between the years 1977 and 2000 inclusive. The main purpose of analyzing this data set is to give a larger multigene example and to compare across-site rate heterogeneity inferred here with the previous study. Therefore, we do not report results for simpler models as we do for the single-locus data sets.

Dirichlet Process Priors

To complete our SDPM1 and SDPM2 construction, we need to specify base distributions for the Dirchlet process(es). When specified hierarchically (Suchard et al. 2003), these distributions allow for the sharing of information across random partitions and the borrowing of strength in parameter estimation. We construct the base distribution for substitution model parameters as $G_0^{\phi}(\boldsymbol{\phi}_k) = G_0^{\rho}(\boldsymbol{\rho}_k)G_0^{f}(\boldsymbol{f}_k)G_0^{\xi}(\xi)$. We use a multivariate normal distribution as the base for $\boldsymbol{\rho}_k, G_0^{\rho}(\boldsymbol{\rho}_k) = \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To induce a hierarchy, mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ are treated as random parameters, where $\boldsymbol{\mu}$ is assumed to have a multivariate normal prior with fixed mean $\boldsymbol{\mu_0}$ and variance $\boldsymbol{\Sigma_0}$. The precision $\boldsymbol{\Sigma}^{-1}$ carries a Wishart prior, with scale matrix $\boldsymbol{V}$ and degrees of freedom d.

We constructed informative priors on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the analyses on the RNA virus data sets according to the following procedure. We analyzed 26 RNA virus data sets (listed in supplementary table S1, Supplementary Material online) from Jenkins et al. (2002) with GTR + Γ_4 using (Guindon et al. 2010, Phyml). Γ_4 models the rate across site with discretized gamma distribution with four categories. The maximum likelihood estimates (MLEs) of the relative rates in the GTR model were transformed to the space of $\boldsymbol{\rho}_k$. Using the mclust package in R (Fraley and Raftery 2002, 2006), we fitted a multivariate normal distribution to these estimates across the data sets, yielding $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma_0}$. There is little information on how the variance $\boldsymbol{\Sigma}$ should vary across sites, so we set $\boldsymbol{V} = \boldsymbol{\Sigma_0^{-1}}$ and $d = 7$, so that the prior mean of $\boldsymbol{\Sigma}$ matches $\boldsymbol{\Sigma_0}$. Informative priors on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for analyses on the mammal data set were also constructed according to the procedure above with 25 mammal data sets (listed in supplementary table S2, Supplementary Material online) randomly selected from Ranwez et al. (2007).

The base distribution of nucleotide base frequencies G_0^f is formulated as follows:

$$\begin{aligned} G_0^f(\cdot) &= \text{Dirichlet}(\eta \times \boldsymbol{q}), \\ \boldsymbol{q} &\sim \text{Dirichlet}(1, 1, 1, 1), \\ \eta &\sim \text{Gamma}(0.001, 0.001), \end{aligned} \tag{12}$$

where η is the dispersion parameter and $\boldsymbol{q} = (q_A, q_C, q_G, q_T)$ is the across-partition mean frequencies. The base distribution of the substitution model indicator G_0^{ξ} is given by

$$\begin{aligned} G_0^{\xi}(\cdot) &= \text{Multinomial}(\boldsymbol{p}), \\ \boldsymbol{p} &\sim \text{Dirichlet}(1, \ldots, 1), \end{aligned} \tag{13}$$

where $\boldsymbol{p} = (p_{K80}, \ldots, p_{GTR})$ are the across-partition model probabilities. Having these hierarchical prior parameters $\boldsymbol{q}$, η and $\boldsymbol{p}$ will improve mixing for the partition allocation variables. The parameterization of our Q matrix can also accommodate Jukes et al. (1969, JC69). However, this set up of mixture model treats categories with ξ = JC69 having different ρ and/or f values as different categories. This is not preferable as these categories have effectively the same model. Therefore, we exclude JC69 from our model to avoid this problem.

The base distribution of rate $G_0(r)$ is assumed to be a lognormal distribution and takes the form

$$\begin{aligned} G_0(\log r_k) &= \text{Normal}(\zeta, \sigma^2), \\ \zeta &\sim \text{Normal}(\mu_\zeta, \sigma_\zeta^2), \\ \sigma^{-2} &\sim \text{Gamma}(\alpha_{\sigma^2}, \beta_{\sigma^2}), \end{aligned} \tag{14}$$

where ζ is mean and σ^2 is the variance.

For the analyses on the serially sampled RNA virus data sets (EBOV and RSVA), informative prior on ζ is constructed by fitting a lognormal distribution (Venables and Ripley 2002) to the MLEs of substitution rate across 50 data sets presented in Jenkins et al. (2002). The log-space mean and standard deviation of the fitted lognormal distribution are assigned to μ_ζ and σ_ζ^2, respectively.

In analyses of contemporaneous sequences (like the mammal and HCV-4 data sets), rate and time cannot be separated without node calibrations. Usually, one would fix the rate to 1.0 and estimate the tree height in substitution units. As our DPM models estimate the rate multipliers, ideally we would like to fix the tree height to 1.0. However, doing so forbids some proposal moves that are important for efficient traversal of tree space. Therefore, we use a narrow normal prior, Normal(1.0, 0.1), on the tree height. This reduces the problem of nonidentifiability and permits useful tree proposals. We assume the log-space mean of the rate base distribution, ζ, is from Normal(−2.3, 2.35). Thus, the median of the base distribution, e^{ζ}, is assumed to come from Lognormal(−2.3, 2.35). This lognormal distribution has 2.5%, 50.0% (median), and 97.5% quantiles of 0.001, 0.1, and 10.0, respectively. It is a broad prior that covers the range of relevant tree heights (measured in substitutions per site).

The gamma prior applied to σ^{-2} has shape $\alpha_{\sigma^2} = 1$ and rate $\beta_{\sigma^2} = 0.1$, which is a fairly broad exponential distribution with variance of 100.

Following the analyses presented in model selection method articles of Lemey et al. (2009) and Heled and Drummond (2010), we also place 50% prior probability on the most parsimonious model by setting the χ of SDPM1 and $\chi(\Phi)$ and $\chi(r)$ of SDPM2 to values, such that the prior probability is 0.5 for $K = 1$ for SDPM1 and K^{Φ} and K^{r} for SDPM2.

Analysis

The data sets were analyzed with HKY + Γ_4 + I, GTR + Γ_4 + I, SRD2006 (GTR + Γ_4 + I for

each codon position), GY94 + Γ_4 + I, SDPM1, and SDPM2. In addition, the data sets were also analyzed using SDPM2 with K^{Φ} fixed to 1. This special case of the SDPM2 is labeled RDPM (rate Dirichlet mixture model), which is very similar to the model presented by Huelsenbeck and Suchard (2007). The rates across sites are not normalized when using RDPM, SDPM1, or SDPM2.

For each data set and substitution model, we analyze them with a strict clock model and an uncorrelated lognormal relaxed molecular clock (Drummond et al. 2006, LNRC). To extract the absolute site rates (or site tree heights if calibration is absent), the branch rates are normalized to 1.0.

Analyses of all virus data sets used a Bayesian skyline plot coalescent prior (Drummond et al. 2005), whereas the Mammal data set had a Yule process prior.

The first 10% steps of the MCMC are discarded as burn-in. The convergence and quality of mixing was examined by using Tracer v1.5 (Rambaut and Drummond 2009). Supplementary table S3, Supplementary Material online, presents the MCMC chain lengths for each analysis. The marginal likelihood of each analysis was approximated using the method proposed by Newton and Raftery (1994) with the stabilization made by Redelings and Suchard (2005).

All input XML files for the analyses performed and the source code for the BEAST 2 add-on that implements the described methodology are available from http://code.google.com/p/subst-bma/(last accessed December 6, 2012). This add-on consists of 1) priors for model parameters, 2) a suite of proposal moves for sampling the partition via Gibbs and Metropolis–Hastings sampling, 3) extensions to likelihood calculations, and 4) components that enable BEAST 2 to handle a variable number of models during the MCMC.

To infer the posterior distribution of the tree topology, we use a series of proposal moves, including narrow exchange, wide exchange (Drummond et al. 2002), Wilson–Balding (Wilson and Balding 1998), and subtree-slide. Subtree-slide is similar to moves proposed by the LOCAL operator (Mau and Newton 1997; Mau et al. 1999; Larget and Simon 1999). Details of these moves are described in Höhna et al. (2008) and have been implemented in both the BEAST 1 and BEAST 2 software packages.

Simulation Study

Simulated data sets are generated under two procedures. In the first procedure, we randomly drew parameters of a GTR model and the shape parameter α of a Gamma-distributed site rate model from empirically derived distributions fit to the 25 virus data sets as described in the Dirichlet process prior section. We then drew four site-specific rate values from a Gamma distribution with shape set to α. Each site in the alignment was assigned to one of the four rates with equal probability. Using the randomly drawn GTR model and site rates, sequences were simulated on a tree with 30 taxa randomly drawn from a Yule model with a birth rate of 20. Here, the true value of $K^{\Phi} = 1$ and $K^{r} = 4$. One hundred data sets were simulated under this procedure, and each of them is analyzed with RDPM, SDPM1, and SDPM2.

In the second procedure, we randomly drew 100 sets of model partitions and tree from posterior of the HCV-4 data set analyzed with SDPM2 and strict clock model. Sequences were simulated with 411 sites. These data sets are analyzed with SDPM2.

The priors on the hyperparameters of Dirichlet process base measure are the same as those used for analyzing HCV-4 data set. In all simulation analyses, we fixed the concentration parameter to the value that gave rise to prior probability of 0.5 for $K/K^{\Phi}/K^{r} = 1$. We then repeated all the simulation analyses but allowed the concentration parameter to be estimated. We assumed

$$\chi \sim \text{Exponential}(\gamma), \tag{15}$$

where the rate, γ, was set to a value, such that the prior probability is 0.5 for $K/K^{\Phi}/K^{r} = 1$. γ therefore was set to 0.135 for the simulated sequences with 1,000 sites and 0.154 for those with 411 sites.

Results

Heterogeneity in Substitution Patterns

The posterior distributions of the number of category parameters K, K^{Φ}, and K^{r} provide some indication of the level of heterogeneity in the substitution process across sites. Figure 1 presents the posterior distributions of K, K^{Φ}, and K^{r}, as well as their prior distribution in each mixture model analysis. Although each of K, K^{Φ}, and K^{r} takes the value 1 with prior probability of 0.5, most analyses exclude $K = 1$ when analyzed with SDPM1 and exclude $K^{\Phi} = 1$ and $K^{r} = 1$ when analyzed with SDPM2 from their respective 95% highest posterior density (HPD) intervals, providing strong evidence for heterogeneity of substitution pattern and rates across sites. The only exceptions are the K^{Φ} estimates for the RSVA data set. The Bayes factor for across-site homogeneity versus heterogeneity of substitution patterns is given by

$$\frac{\text{Posterior } \mathbb{P}\left(K^{\Phi} = 1\right)}{\text{Posterior } \mathbb{P}(K^{\Phi} > 1)} \times \frac{\text{Prior } \mathbb{P}\left(K^{\Phi} > 1\right)}{\text{Prior } \mathbb{P}(K^{\Phi} = 1)}. \tag{16}$$

For RSVA, the Bayes factor is 0.140 for the strict clock analysis and 0.175 for the relaxed clock analysis. While far from definitive, these Bayes factors provide substantial evidence against across-site homogeneity in substitution pattern according to the interpretation scale provided by Jeffreys (1998). A more conclusive outcome may be obtained by adding more sequences. Conditioned on the data set and clock model, the estimated posterior means of K^{Φ} and K^{r} are smaller than that of K, which suggests that less categories of substitution pattern are required if the site rate heterogeneity is modeled separately. However, there is one exception—for EBOV, the posterior mean of K^{Φ} is not smaller than that of K (supplementary table S4, Supplementary Material online).

One question of interest is "Should every site in an alignment be modeled by the same type of nucleotide substitution model?" If not, it is important to infer which substitution model should be used at each site. We present the answer obtained from the DPM model analyses in figure 2, which

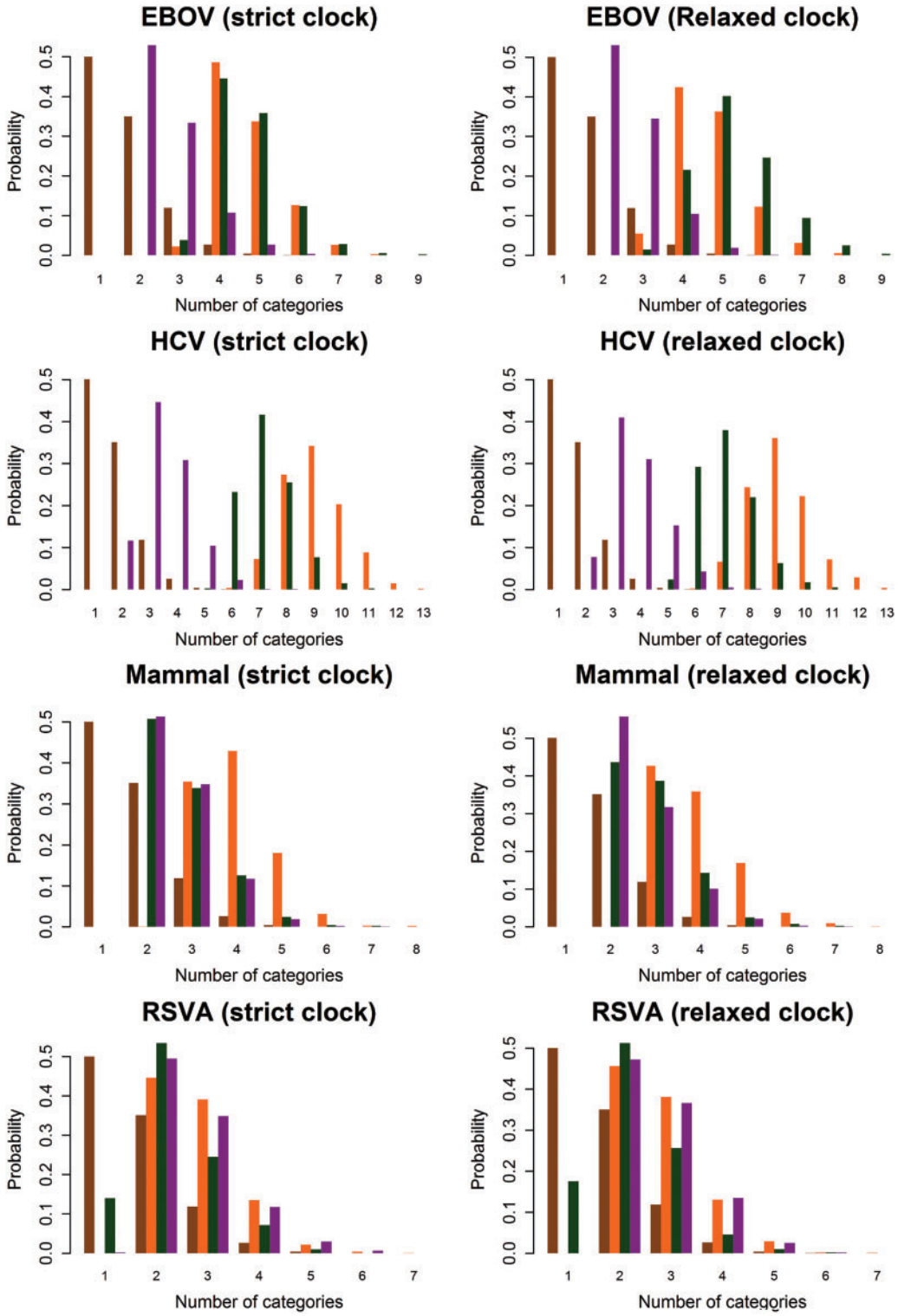

Fig. 1. Posterior distributions of the number of categories for substitution pattern and rates from analyses with mixture models. The prior distribution of the number of categories is in brown. The posterior distribution of K estimated using SDPM1 is in orange. The posterior distributions of K^{Φ} and K^{r} estimated using SDPM2 are colored in green and purple, respectively.

consists of 16 grid plots. In each grid plot, each row represents one of the five nucleotide substitution models and each column represents a site in an alignment. A grid located in row M and column i is colored according to the posterior probability of site i being generated by model M. The color darkens as the probability increases. The posterior average number of sites that have selected an M model is on the right side of the plot. Given a data set and an SDPM model, little difference is seen in the across-site substitution pattern between strict clock and LNRC analyses. However, there appears to be some differences between SDPM1 and SDPM2 analyses. In the SDPM1 analyses on EBOV, there seems to be

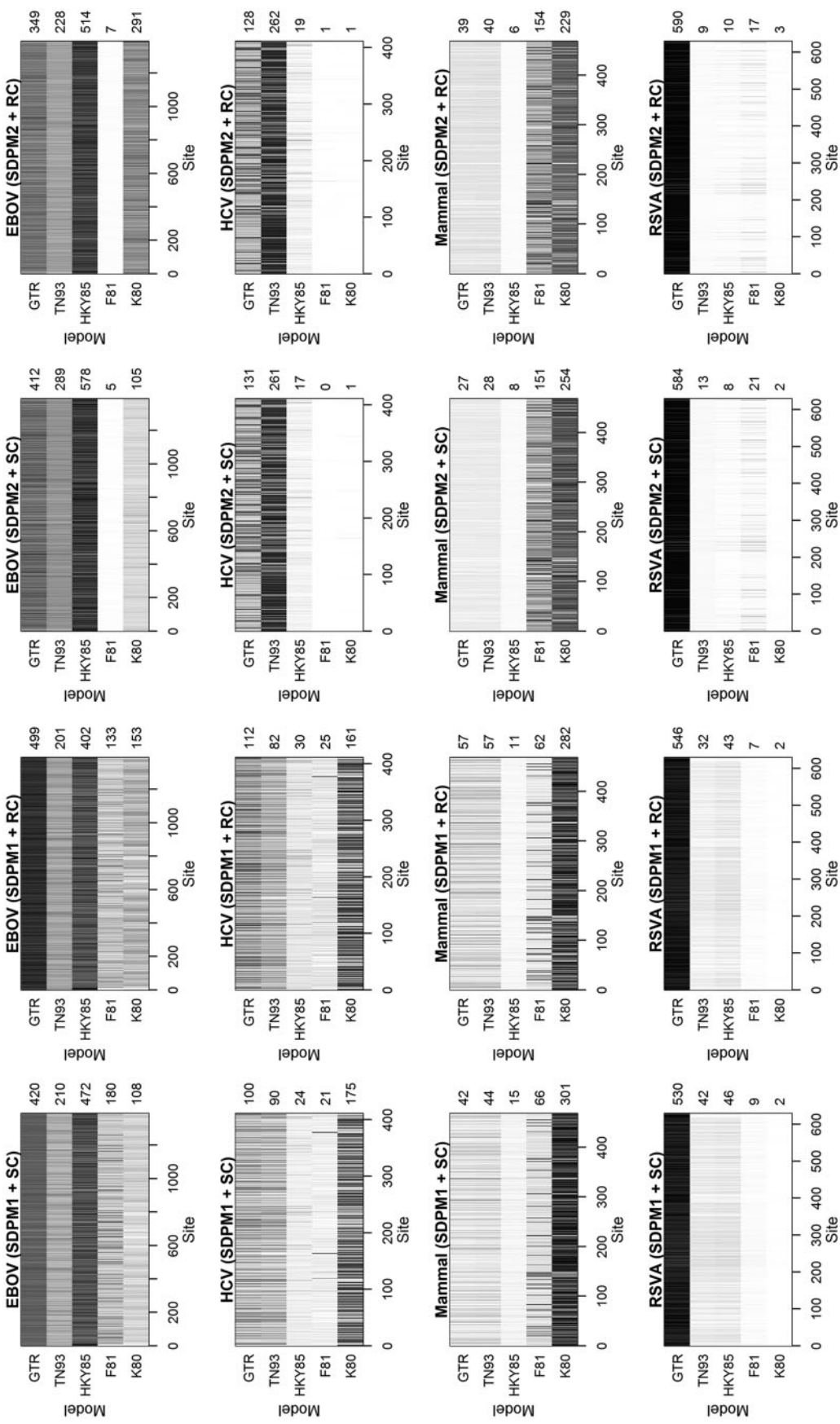

FIG. 2. Support for each model at each site indicated by the posterior probability that a model is selected to fit that site. The color becomes darker as the posterior probability increases. The average number of sites fitted by a model is indicated on the axes on the right hand side of the plots.

some support for F81; however, this is not evident after switching to SDPM2 as illustrated by the white band in the F81 row. An even larger contrast is displayed by the analyses on HCV-4. The results from the SDPM1 analyses on HCV-4 suggest that the most favored model is K80; however, the SDPM2 analyses show almost no support for K80 and clear preference for TN93 and GTR. All SDPM analyses on Mammal prefer K80, but this seems stronger in the SDPM1 analyses. In contrast, the reverse pattern is observed in the analyses on RSVA, where all analyses prefer GTR, but the preference is stronger in the SDPM2 analyses.

Figure 2 does not provide information on whether two sites *i* and *j*, which both prefer a specific type of model, are modeled by the same parameter values. That is, if site *i* prefers a GTR model it does not follow that site, *j* also prefers the same GTR parameter values. To illustrate the cluster structure, we performed cluster analyses on the estimates of substitution model parameters using *k*-means algorithm implemented in the R package MASS (Venables and Ripley 2002; R Development Core Team 2011). Let K_{max}, K^{Φ}_{max}, and K^{r}_{max} represent estimated posterior mode of K, K^{Φ}, and K^{r}, respectively. The number of clusters is predefined in the *k*-means algorithm. Cluster analyses on SDPM1 parameter estimates have K_{max} clusters, whereas those on SDPM2 parameter estimates have K^{Φ}_{max}. As examples, we present the results from the cluster analyses for the mammal (fig. 3) and RSVA (fig. 4). Figure 3 shows that sites are indeed clustered according to the model most preferred. Those that have chosen K80 tend to be in one cluster, and those prefer F81 is in another cluster. This segregation does not appear in the results for RSVA (fig. 4). Although most sites prefer the GTR model, there is still grouping structure, in other words, they are not modeled by the same GTR.

Because all data sets used in this study code for proteins, we would like to see whether the across-site heterogeneity in rate uncovered by our mixture models corresponds to codon positions. For each MCMC step that has K_{max} categories, we first order the categories in increasing order of the rate, so that category 1 has the slowest rate, whereas category K_{max} has the fastest rate. The proportion of sites in each category is computed for each codon position. The same procedure is repeated for the results from SDPM2 analyses, except K_{max} is replaced by K^{r}_{max}, the number of rate categories with the highest posterior probability. Figure 5 illustrates the posterior mean proportions of sites in category 1 to category K_{max} for every SDPM1 analysis and the posterior mean proportions in category 1 to K^{r}_{max} for every SDPM2 analysis. The bars are colored according to the proportion of sites in each category, and the category with a faster rate is closer to the top of the bar. All analyses show that in general the third codon position has a higher substitution rate, although there is much variation within the codon positions. This increase in the third codon rate is concordant with previous findings (Huelsenbeck and Suchard 2007).

We examine whether the preference for the type of substitution model also differs by codon position. For each state, we compute the proportion of sites at each codon position selecting each one of the five types of substitution model. The posterior mean proportions for each codon position are presented in the plots shown in supplementary figure S1, Supplementary Material online. SDPM1 analyses show that the preference for the type of substitution model seems to differ by codon positions. For EBOV, HCV-4, and Mammal, sites in the third codon position appear to prefer more complex substitution models, but the difference is not so apparent in the RSVA data set. In contrast, the SDPM2 analyses do not show any significant difference in preference for substitution models across codon positions.

We compute the relative standard deviation (RSD) for the substitution model parameter values across the categories. RSD is the standard deviation divided by the absolute value of the mean. The values of posterior mean and 95% credible interval boundaries of RSD are presented in figure 6. Analyses with SDPM1 on EBOV produce relative rate parameters with mean RSD values around 1, except for the rate between C and T. Analyses on HCV-4 and mammal produce mean RSD values around 1 for relative rates, other than that between C and T. These RSD values suggest reasonably clear signal of heterogeneity in substitution pattern, which is likely to have contributed to the difference in model choice across sites as shown in figure 2. All posterior mean RSD values estimated from RSVA are between 0.15 and 0.7, which are generally lower than those produced by other data sets. It suggests that the signal for heterogeneity in substitution patterns is not strong. Moreover, it is consistent with a higher posterior probability for homogeneity in this data set than others.

Model Comparison

The Bayes factor is often used for model comparison in Bayesian analysis, expressing the ratio of the marginal likelihoods of two competing models. The marginal likelihood is the likelihood of the data given the model and is integrated across the entire parameter space of the model. It therefore accounts for the complexity of the model and penalizes greater model complexity. The natural logarithm of the marginal likelihoods of all single-locus analyses are presented in table 2, and their differences are log Bayes factors.

The substitution models are of increasing complexity from left to right. Conditioned on a data-clock model combination, the worst fit to the data is found in nucleotide substitution models that do not account for across-site heterogeneity in substitution patterns and do not estimate rate partitioning. Allowing restricted heterogeneity by performing codon partition substantially improves the marginal likelihood for all data sets except for RSVA. Increasingly flexible partition schemes of the substitution pattern improve the fit of the model substantially. This outcome indicates that codon partitioning does not fully characterize the complexities of across site variation in protein coding sequence alignments.

The fit of $GY94 + \Gamma_4 + I$ relative to other models varies considerably across different data sets. For the Mammal data set, $GY94 + \Gamma_4 + I$ fits the data just, as well as the SDPM models. Similarly, the SDPM models do not fit EBOV substantially better than the codon model. The detected heterogeneity of these two data sets may therefore be just as easily

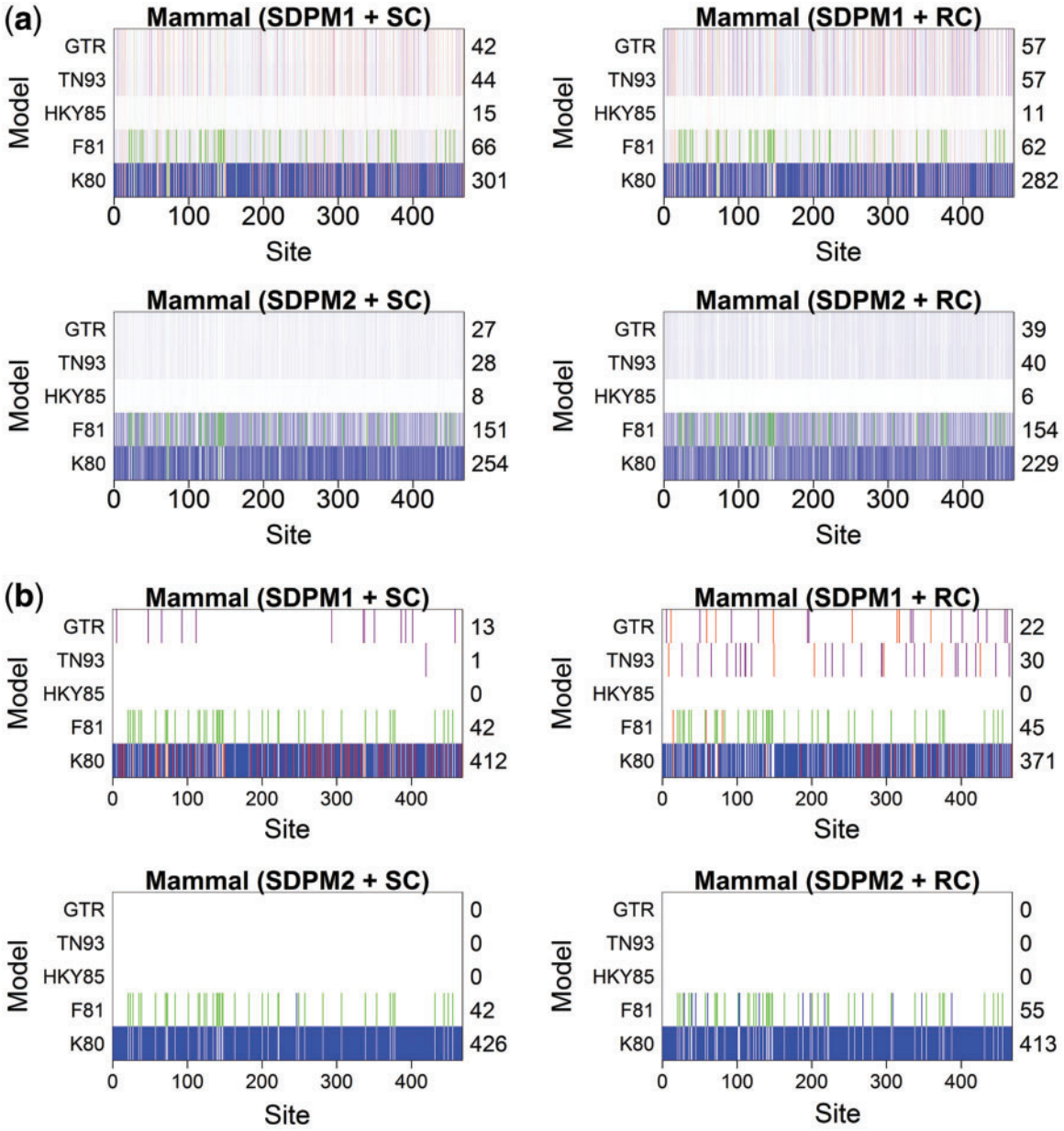

FIG. 3. Support for substitution models at each site of the Mammal data set, with each site colored according to the cluster to which it is assigned by the cluster analysis performed on the estimates of substitution model parameters. For the SDPM1 analyses, the sites are grouped into four clusters as the marginal posterior probability of K is the largest when $K = 4$ and the colors used to distinguish them are blue, green, purple, and orange. For the SDPM2 analyses, $K^{\Phi} = 2$ has the largest marginal posterior probability, so the sites are grouped into two clusters colored with blue and orange. The posterior probability is indicated by the darkness of the color in part (*a*). Darker coloring corresponds to higher probability. Only the model with the highest posterior probability (best model) at each site is colored in part (*b*), and the number of sites that selects a model as the best model is reported on the axis on the right hand side.

explained by a simple codon-based model. However, the codon model does not fit the RSVA and HCV-4 data sets, as well as the SDPM models. For those two data sets, the SDPM models have substantially better marginal likelihoods than all the other substitution models. This suggests that the heterogeneity in these two protein coding sequences cannot be fully explained by the genetic code or at least the properties of the genetic code incorporated in codon model tested here.

For the data sets HCV-4, Mammal, and RSVA, the difference in the marginal likelihood between SDPM1 and SDPM2 is < 50 natural log units. However, for EBOV, the difference is > 150 natural log units and the log marginal likelihood difference between SDPM1 and SDPM2 is 16–19 times the difference between HKY and GTR. Therefore, the improvement in model fit of SDPM2 over SDPM1 can sometimes be very substantial.

Estimation of Phylogenetic Parameters and Their Hyperparameters

The tree height estimates are shown in figure 7. Given a clock model, the mixture models tend to produce older trees than

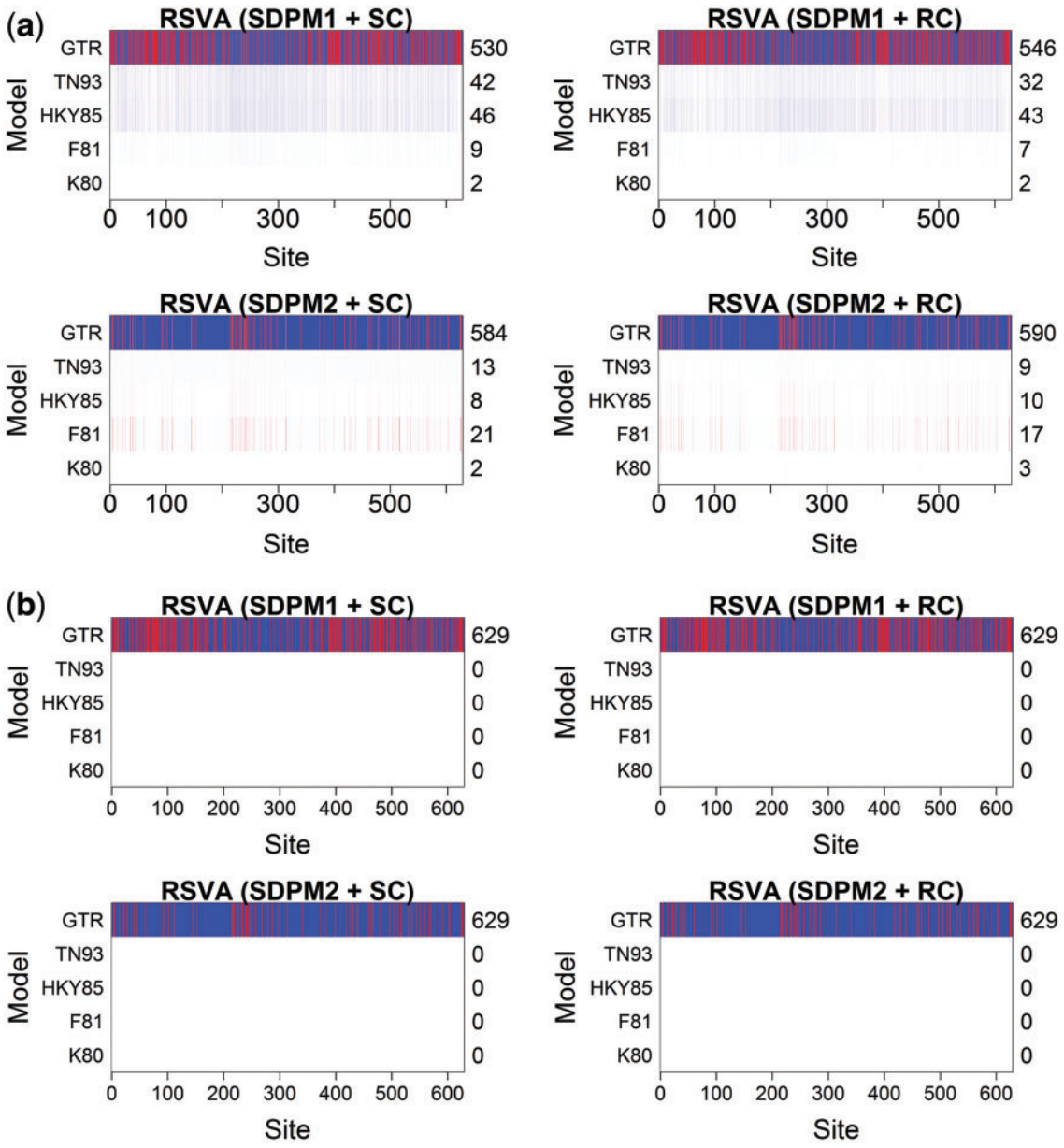

FIG. 4. Support for substitution models at each site of the RSVA data set, with each site colored according to the cluster to which it is assigned by the cluster analysis performed on the estimates of substitution model parameters. The posterior mode values of K and K^{Φ} are equal to two; therefore, for both SDPM1 and SDPM2 analyses, sites are grouped into two clusters colored blue and red. The posterior probability is indicated by the darkness of the color in part (*a*). Darker coloring corresponds to higher probability. Only the model with the highest posterior probability (best model) at each site is colored in part (*b*), and the number of sites that selects a model as the best model is reported on the axis on the right hand side.

other simpler substitution model partitions for EBOV. The estimated posterior means of the EBOV tree height under SDPM models are between 51% and 61% older than that of other nucleotide models for strict clock analyses and are between 40% and 62% older for relaxed clock analyses. The codon model analyses and SDPM analyses have similar tree height estimates. The results from the strict clock analyses on HCV-4 show that the tree height estimates of SDPM models are, in contrast, 34–52% shorter than that of other models. Moreover, the SDPM models produced even shorter trees (68–78% shorter) in LNRC analyses than in strict clock analysis. However, the difference in tree length estimates is much smaller between DPM models and others. The posterior mean tree length is between 3.53 and 3.97 for SDPM models and 4.41 and 4.83 for non-SDPM models. This suggests that the SDPM models only reduced the lengths of a few branches in the trees near the root. The analysis with the GY94 + Γ_4 + I model produced a much taller Mammal tree than all nucleotide substitution models, among which the tree height estimates do not display substantial differences. For the RSVA data set, the tree height estimates do not vary significantly across all substitution models given a strict clock model.

To ease tree-space visualization, we have subsampled 100 trees from each posterior tree distribution. For the 700 trees obtained from the same clock model and data set, we

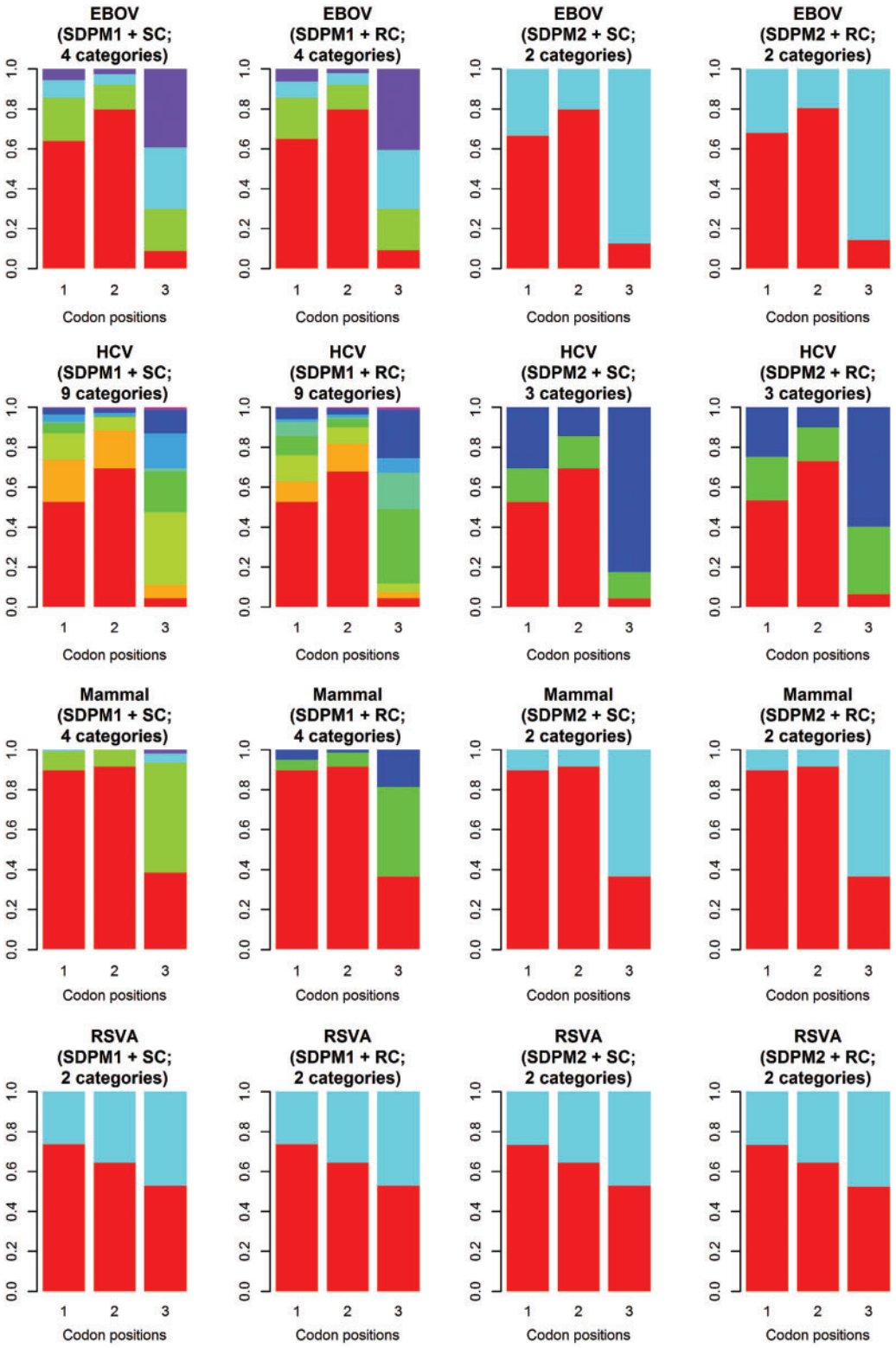

FIG. 5. Proportion of sites in each codon position as a function of rate. The number of category shown has the maximum posterior probability. Each bar represents a codon position, and it is colored according to the posterior mean proportion of sites in each rate category. The colors are picked from the "rainbow" scheme, and clusters with faster mean rate are in colors closer to the violet end.

compute the Robinson-Foulds distance between each tree. We apply principle coordinate analysis (PCO) on the 700×700 distance matrices. Supplementary figure S2, Supplementary Material online, presents the reduced-space plots with the scores on the first two major principle axes. Each point represents a tree from the subsample. Of the four data sets, only the posterior distributions of HCV-4 produced reduced-space plots that displayed clustering by site model (each model was distinguished by a different color) (fig. 8). There appears to be three major groupings by model:

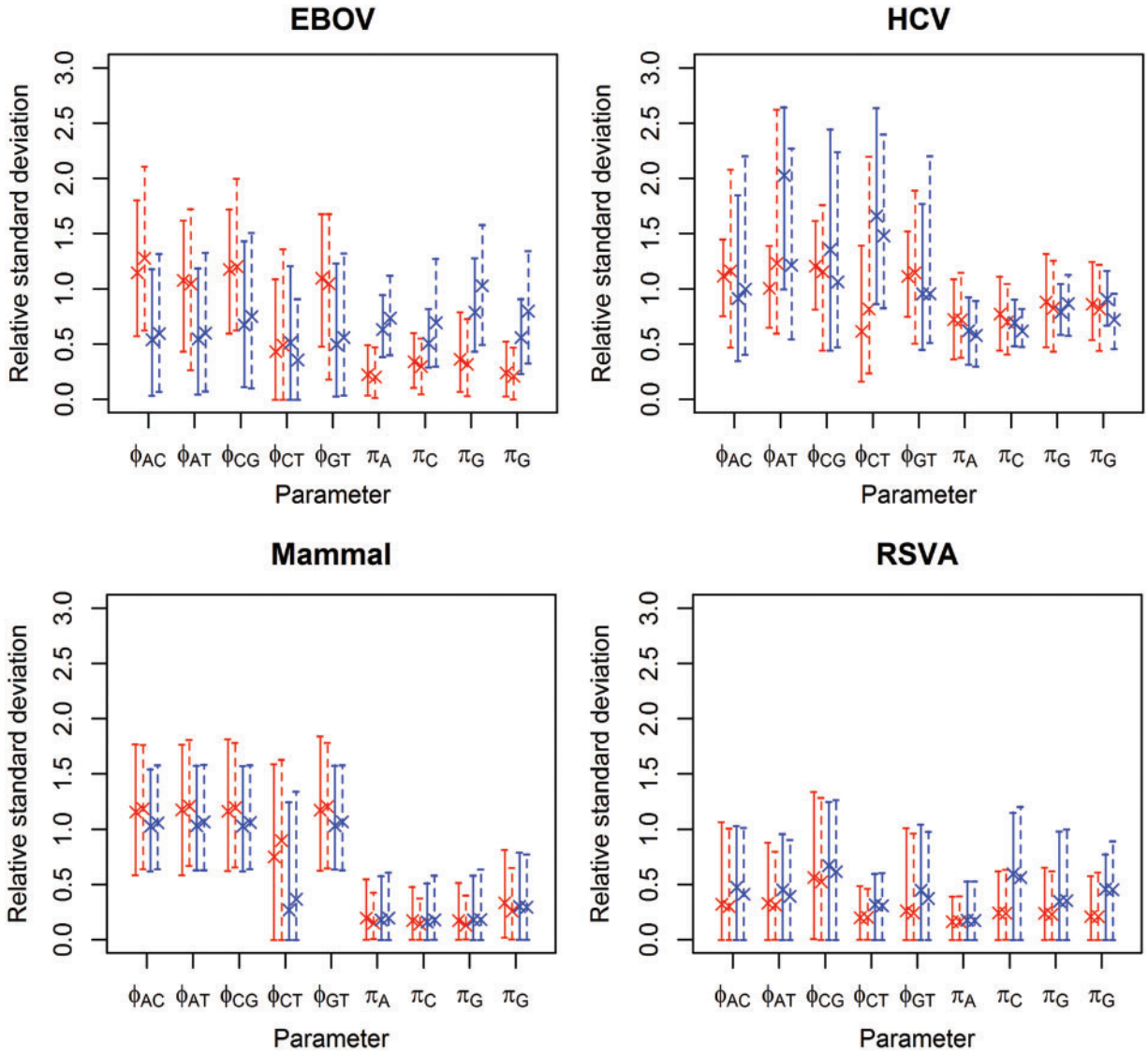

FIG. 6. Posterior RSD of substitution model parameter values across categories. Analyzed with SDPM1 are in red, SDPM2 in blue, strict clock model in solid lines, and lognormal relaxed clock model in dotted lines.

Table 2. The Natural Log Marginal Likelihoods of Analyses with Strict Clock Model.

Data Set	Clock Model	HKY+Γ_4+I	GTR+Γ_4+I	SRD2006	GY94+Γ_4+I	RDPM	SDPM1	SDPM2
EBOV	SC	−7,495	−7,487	−7,114	−6,734	−6,914	−6,682	−6,531
EBOV	LNRC	−7,479	−7,468	−7,093	−6,714	−6,892	−6,648	−6,475
HCV-4	SC	−6,172	−6,167	−6,041	−6,208	−5,860	−5,638	−5,601
HCV-4	LNRC	−6,153	−6,147	−6,017	−6,190	−5,814	−5,596	−5,550
Mammal	SC	−1,695	−1,689	−1,582	−1,522	−1,570	−1,534	−1,517
Mammal	LNRC	−1,690	−1,681	−1,578	−1,518	−1,565	−1,523	−1,511
RSVA	SC	−3,112	−3,093	−3,072	−3,132	−2,995	−2,988	−2,979
RSVA	LNRC	−3,108	−3,091	−3,068	−3,130	−2,987	−2,987	−2,976

GY94 + Γ + I (green) stands out as a single model; the SDPM1 (blue) and SDPM2 (purple) seem clearly separated from the common nucleotide substitution models, HKY + Γ + I (red), GTR + Γ + I (orange), and SRD2006 (yellow). RDPM (turquoise) scatters between SDPMs and the common nucleotide substitution models. It is natural that RDPM bridges the two groupings as it does not partition the alignment for substitution models, but it does estimate the substitution model and across-site rate variation with a DPP.

To further investigate the differences in tree topology of HCV-4, we record all the unique clades and their posterior

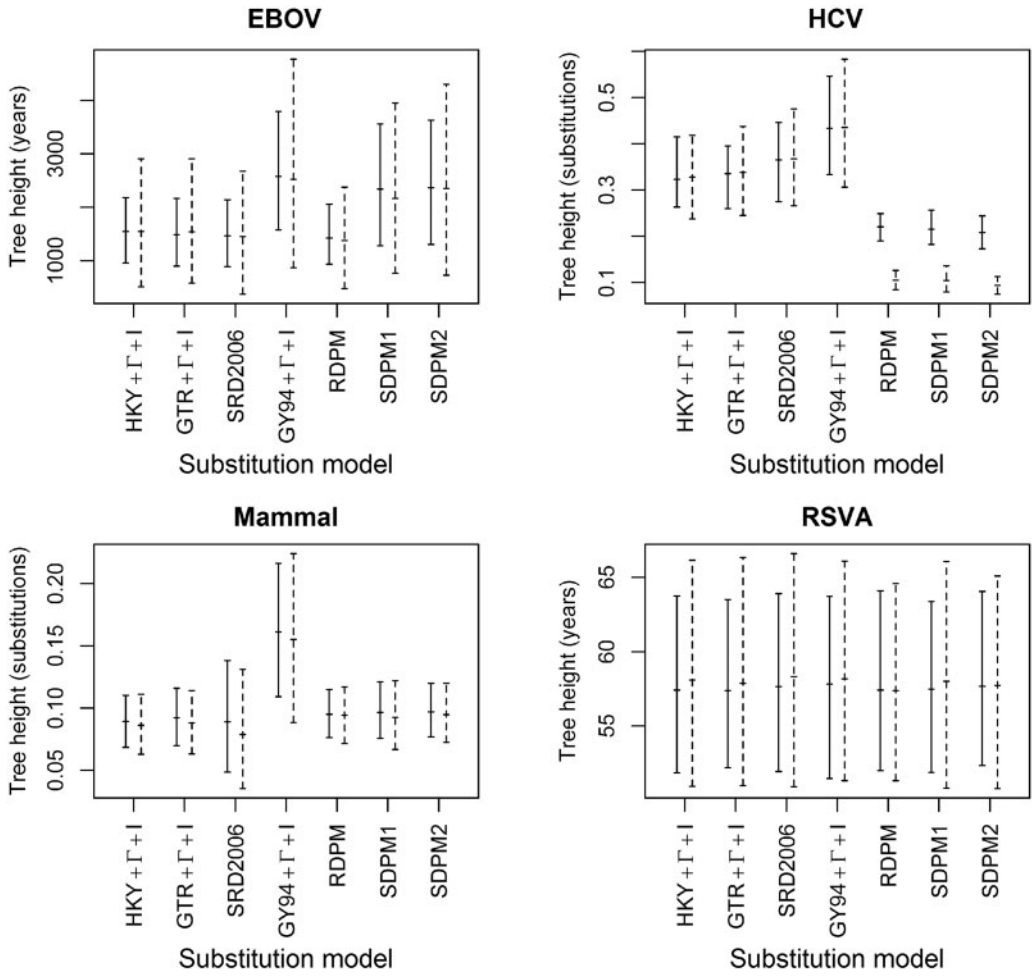

FIG. 7. Tree height estimates. Each bar spans the 95% HPD of the tree height, and the posterior mean is marked on the bar. Solid bars are estimates from strick clock analyses, whereas the dashed bars are estimated from the lognormal relaxed clock analyses.

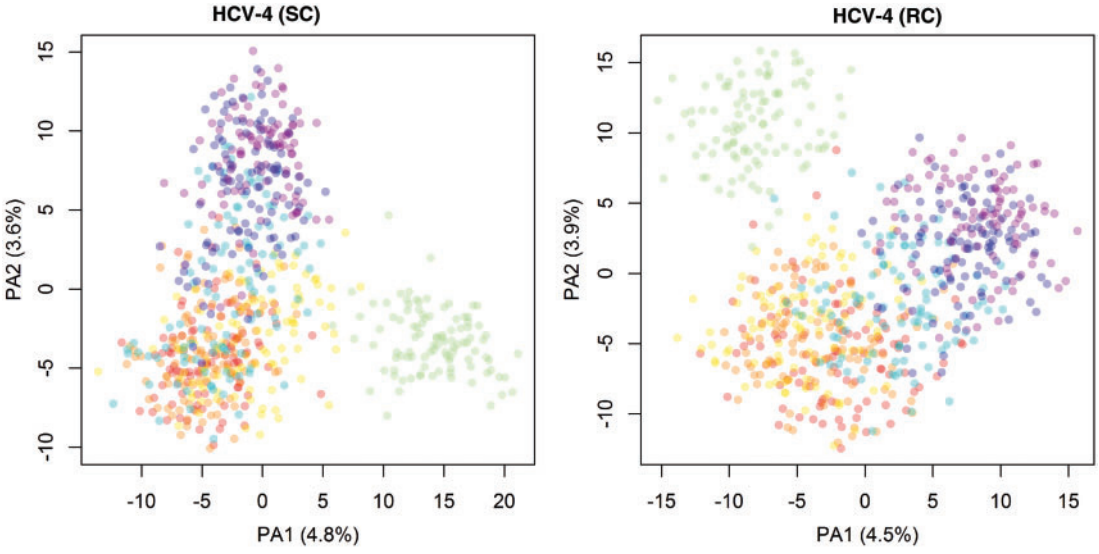

FIG. 8. Reduced space of substitution models based on clade posterior probability estimated from HCV-4. Each point represents a tree from the subsample. The trees are colored according substitution model used in the analysis. HKY+Γ+I is colored red, GTR+Γ+I orange, SRD2006 yellow, GY94+Γ+I green, RDPM turquoise, SDPM1 blue, and SDPM2 purple.

probability in each of the two posterior tree distributions. Conditioned on a clock model, each substitution model has a vector of posterior probabilities for each clade. We use clade posterior probabilities to find the Manhattan distance between each pair of substitution model parameters. A 7×7 distance matrix is constructed for the substitution models. A PCO analysis is performed on this distance matrix, and the reduced-space plots with the first two major PAs are presented in figure 9.

The same groupings appear again in these plots. For each clade, we find the range (max−min) of posterior probabilities across the seven substitution models. The top 50 clades with the highest range of posterior probability have range values between 0.278 and 0.882 for strict clock analysis and between 0.258 and 0.793 for relaxed clock analysis. Difference in clade support indicates that different substitution models support different topologies. We select GTR + Γ + I, GY94 + Γ + I, and SDPM2 as representatives of each cluster. The top 50 clades with the highest range of posterior probability are mapped to the maximum clade credibility trees of HCV-4 produced by those substitution models (supplementary figs. S3–S8, Supplementary Material online).

To provide some indication on how the posterior distribution on tree topology differs across the different substitution models, supplementary table S5, Supplementary Material online, presents the 95% credible tree sets and the 50% and 5% credible clade sets.

The Bayesian skyline plots for the virus data sets are presented in figure 10. The discrepancies in the tree height estimates of a given data set are reflected in the time frame of the BSPs. For EBOV, the population size estimates produced by the DPM models are much larger at a given time than those produced by other across-site substitution-rate models in both strict clock analyses and relaxed clock analyses. However, all the across-site substitution-rate models shares the same pattern in how population changes over time—they all show that the population of the EBOV is constant up to approximately 100 years ago followed by a bottleneck. The population size estimates and time frame have been rescaled for the results on HCV-4 by using a previously estimated substitution rate 7.9×10^{-4} (Pybus et al. 2001). The BSPs from the strict clock analyses shows that population sizes are quite similar across all substitution models. This suggests that the population size of HCV-4 in Egypt was constant until a rapid expansion occurred approximately 60 years before sample collection. However, the LNRC analyses with the mixture models on HCV-4 suggest a slightly earlier expansion date than other relaxed clock analyses. Given a strict clock model, BSPs estimated for RSVA are very similar across all across-site substitution-rate models.

The 95% HPD intervals and the estimated posterior mean of the birth rate of the Yule process prior are very similar across all analyses with nucletode substitution models on Mammal. The lower bound the 95% HPD interval is between 9.22 and 11.48, whereas the upper bound is between 33.34 and 38.29. The posterior mean ranges from 20.56 to 22.38. This indicates that the inference on birth rate is not affected by the choice of nucleotide substitution model in this case. Birth rate estimates inferred from GY94 + Γ_4 + I are much lower. The strict clock analysis estimates a posterior mean (95% HPD interval) of 14.6 (5.87–23.5), which is similar to that inferred from the LNRC analysis 15.0 (6.07–25.7).

Hepatitis C Virus Subtype 1b Full-Genome Data

Figure 11 displays the 95% HPD intervals of site-specific rates from the RDPM + LNRC analysis on HCV-1b genome sequences. The rest of the results from RDPM and SDPM1 analyses are presented in supplementary figure S9, Supplementary Material online. Comparing with Figure 1(a) from Gray et al. (2011), our results also show a hot spot around 1,250th site, whereas the rate is fairly uniform across the rest of the genome. This is probably why the entire genome (HCV-1b) does not require many more rate

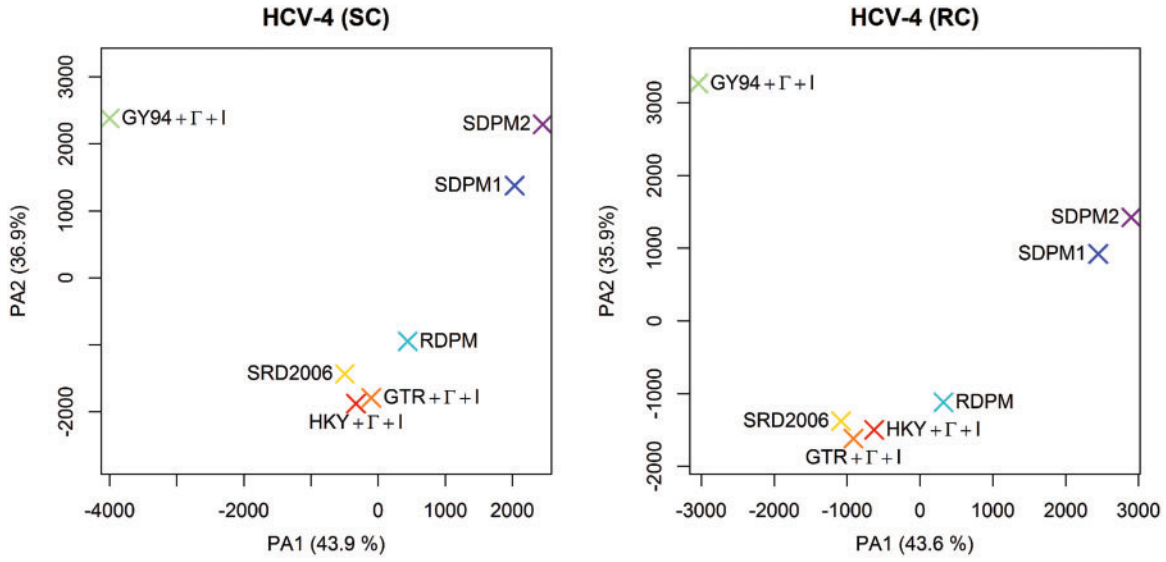

FIG. 9. Reduced space of substitution models based on clade posterior probability.

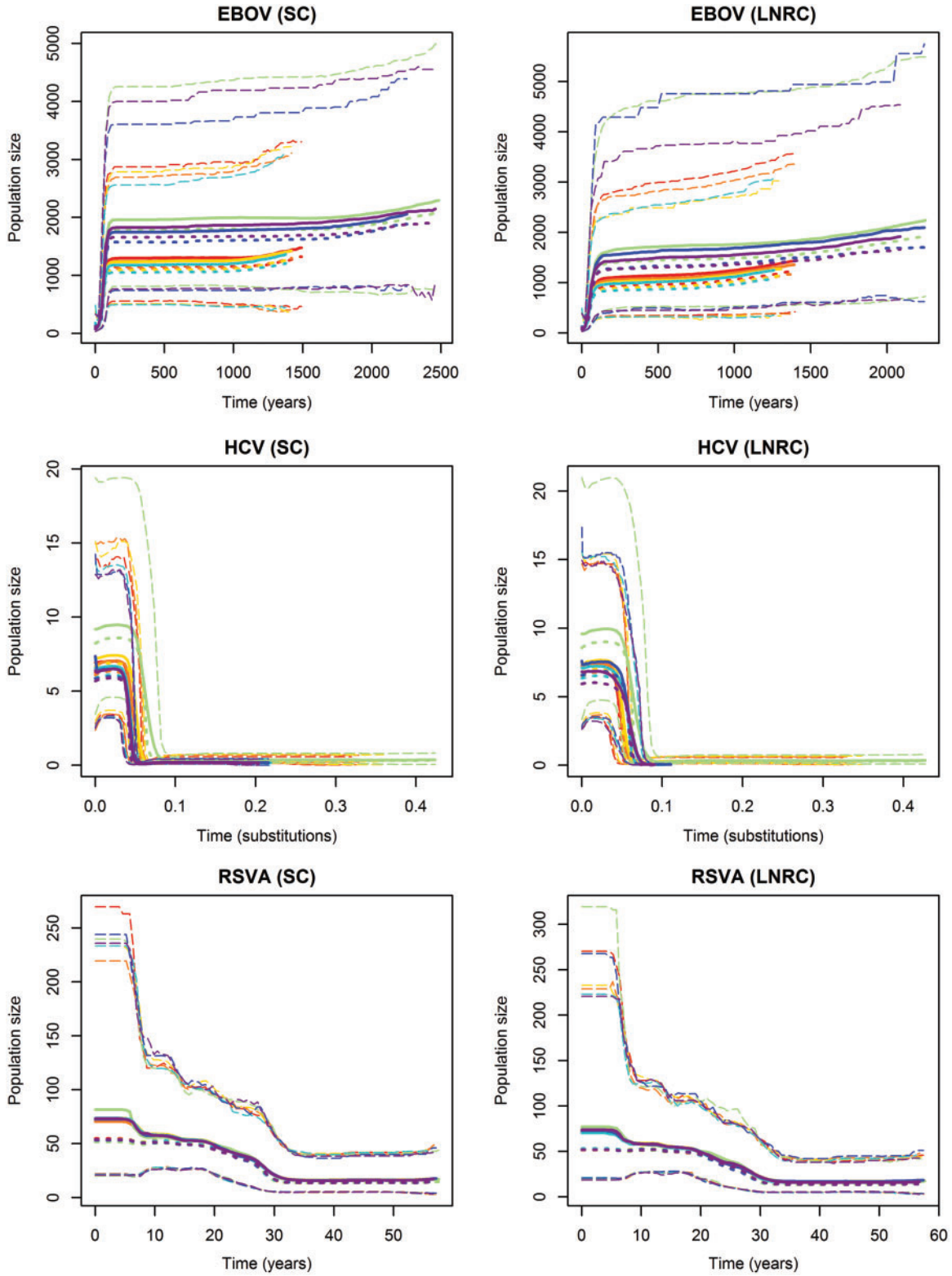

Fig. 10. Bayesian skyline plots for the analyses on EBOV, HCV-4, and RSVA. Each plot presents BSPs estimated under HKY+Γ+I (red), GTR+Γ+I (orange), SRD2006 (yellow), GY94+Γ+I (green), RDPM (turquoise), SDPM1 (blue) and SDPM2 (purple) for a given data set and clock model.

categories (supplementary table S4, Supplementary Material online) than the E1 gene sequences (HCV-4). The region with the unusually fast rates is near the border of genes E1 and E2. The plots also suggest that sites at the third codon position have higher rates (long blue upper tails) than others. In addition, supplementary figure S9, Supplementary Material online, shows less variation in rate estimates inferred from SDPM1 model. This could be due to decreased sensitivity because the SDPM1 model does not allow separation of rate and pattern heterogeneity.

Simulations

Averaged values of statistics used to indicate accuracy and precision of our method are presented in table 3. As measures of accuracy, we use relative bias and the frequency of the true value inside the 95% HPD interval. Relative error and relative size of the 95% HPD interval are used to indicate the level of precision. If a data set is generated with K categories, the relative bias is given by $(\hat{K} - K)/K$ where $\hat{K}$ is the posterior mean of K estimated from a simulated data set. The relative error is the absolute value of the relative bias. If the 95% HPD interval of K has upper (I_U) and lower bounds (I_L), the relative size of 95% HPD interval is defined as $(I_U - I_L)/K$.

For all data sets simulated, we generally underestimated the number of rate categories, which is not surprising as the prior strongly favors homogeneity. However, the negative bias is reduced substantially if we estimate the concentration parameter. This may be attributed to the longer tails of the prior distribution on the number of categories when χ is estimated (supplementary fig. S10, Supplementary Material online). RDPM does not estimate the number of substitution model categories. As for SDPM1, the substitution model and rate share the same category structure. The K^{Φ} estimates from the first set of simulations are naturally positive biased as the true K^{Φ} value is the lower bound (1). The K^{Φ} estimates from the second set of simulations tend to be negatively biased if the concentration parameter value is fixed. If we estimate the concentrating parameter value, then estimates of K^{Φ} seem positively biased with smaller magnitude.

Analyses on data sets simulated from the first procedure yielded high 95% HPD coverage of the true number of categories (0.98–1.00). For data sets simulated from the second procedure, HPD coverage is also high for the true number of categories except for K^{Φ} when the concentration parameter is fixed. This is attributed to the strong negative bias of the estimate, when the true number of categories is large.

For both the number of rate and substitution pattern categories, it appears that the size of relative 95% credible interval is smaller when the value of concentration parameter is fixed than when it is estimated. This outcome is expected as estimating the concentration parameter creates greater uncertainty in the prior on the number of categories.

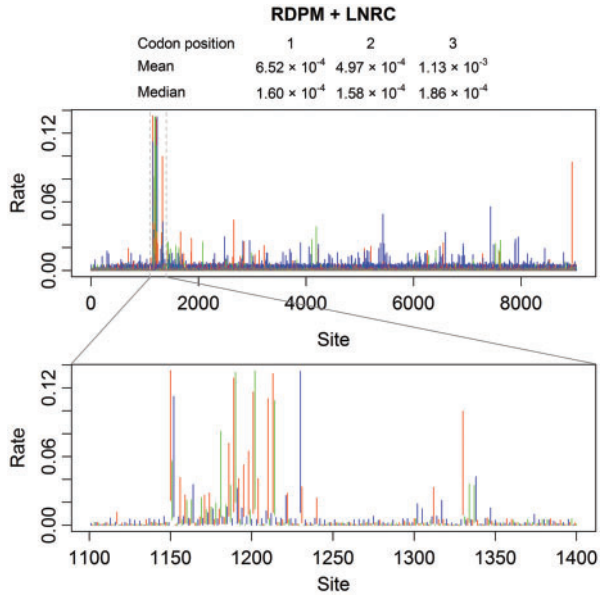

FIG. 11. The 95% HPD intervals of site-specific rates for the HCV-1b genome sequences. Codon positions 1, 2 and 3 are coded in red, green, and blue, respectively.

Discussion

We have presented DPM models that accommodate across-site heterogeneity in both nucleotide substitution pattern and rate. Using Dirichlet process priors enables the estimation of the number of categories required to explain the heterogeneity of nucleotide substitution, as well as the site-to-category assignment. This obviates a priori specification of the partitioning scheme before the analysis. Because the partitioning is carried out at the nucleotide level, our method is more flexible and is not limited to protein coding alignments. More importantly, sites are grouped together based on the similarity of their substitution properties (substitution model or rate parameters) as informed by the data itself.

Table 3. Statistics of Accuracy and Precision of the Estimate of the Number of Categories.

Data Simulation Procedure	Model	Parameter	Estimate χ	Relative Bias	Relative Error	% Inside 95% HPD Interval	Relative 95% HPD Interval Size
One	RDPM	K^r	N	−0.310	0.310	1.00	0.790
			Y	−0.0409	0.122	1.00	1.49
	SDPM1	K	N	−0.306	0.306	1.00	0.810
			Y	−0.0178	0.147	1.00	1.52
	SDPM2	K^{Φ}	N	0.693	0.693	0.98	3.09
			Y	0.885	0.885	1.00	4.23
		K^r	N	−0.307	0.307	0.99	0.795
			Y	−0.0612	0.138	1.00	1.45
Two	SDPM2	K^{Φ}	N	−0.237	0.243	0.61	0.531
			Y	0.140	0.179	1.00	1.15
		K^r	N	−0.138	0.221	0.92	1.14
			Y	−0.105	0.266	1.00	1.81

Similar to previously proposed models that also attempt to accommodate across-site heterogeneity in nucleotide substitution pattern (Huelsenbeck and Nielsen 1999; Pagel and Meade 2004; Shapiro et al. 2006; Whelan 2008), analyses with our DPM models provide evidence supporting the presence of substitution pattern heterogeneity. The SDPM models also reveal that not all sites favor the same type of nucleotide substitution model in our alignment data. These models seem to be able to capture the codon structure in protein coding sequences as evidenced by the tendency to favor faster rate categories in the third codon position. However, it is also clear that there is rate variation among the sites in the same codon position, therefore the pattern of rate variation is more complex than simple codon partitioning.

In some cases, the phylogenetic and hyperparameter estimates produced by the SDPM models are different to those produced by simpler substitution models. For example, the tree height estimates for EBOV produced by the DPM models are substantially older than when using simpler models but similar to that produced by a codon substitution model (Wertheim and Kosakovsky Pond 2011). Perhaps, the heterogeneity found in the data set is the result of selection pressure; however, uncovering the cause of across-site substitution heterogeneity is beyond the scope of this study. The data sets that exhibit significant differences in phylogenetic estimates between DPM model analyses and others also displayed higher levels of across-site heterogeneity in substitution patterns. However, to confirm this trend, a more comprehensive study is required.

The SDPM models fit our four single-locus data sets far better than all standard nucleotide substitution models tested. This is compatible with the presence of across-site heterogeneity of the substitution pattern in the data sets explored. In addition, the large improvement in model fit obtained by SDPM models suggest that simple codon models are not always adequate for protein coding sequences. Our results show that the SDPM models can substantially outperform codon models. As a large prior weight (probability of 0.5) is placed on across-site substitution homogeneity, the variation detected is likely to represent strong evidence of a real signal of site heterogeneity. Because SDPM models can have a large number of parameters (eight free parameters per substitution model category), if the data set is small then overfitting may occur. Overfitting can be prevented by setting the concentration parameter of the Dirichlet process to a smaller value, favoring fewer categories. The substitution model is parameterized, so that the substitution model of each category can be "estimated," achieving site to model assignment. The set of substitution models for selection include models that aim to capture the biological properties observed in nucleotide substitution.

It is quite possible that the most suitable model for a particular (set of) site(s) is not in the set of substitution models we have specified. Fine tuning the set of substitution models may improve the quality of fit. In the model selection study by Huelsenbeck et al. (2004), they have exploited the entire space of 203 possible nucleotide substitution models. Although the most favored models were unnamed ones, in their study they found that the predominant pattern is the difference in the rate between transition and transversion. Moreover, this appears to be the decisive factor for whether or not a model has the highest posterior probability. The models with the highest posterior probability appeared to only have minor difference to named models such as Kimura (1980); Hasegawa et al. (1985). Although most of the favored/best models are unnamed, they still conform to the biological behavior that the standard named models aim to capture/parameterize. Because the differences between the unnamed best model and standard named models are likely to be minor, there should not be drastic differences in the quality of the fit. The relatively small differences in marginal likelihood between HKY and GTR models, when compared with the large differences between them and the SDPM models suggest that modeling improvements that capture rate and pattern heterogeneity across sites will dwarf any gains that might be achieved by providing for intermediate substitution models.

A future improvement of our method is to relax the definition of units of category assignments. Currently, alignment sites are the units of category assignments. If we allow the units to be genes, it may be useful for phylogenomic analyses. In this study, we have not explored the entire substitution model space and have not allowed variation in the topology across partitions. Incorporating either of these properties substantially expands the parameter space, and carefully devised proposal moves would be required to traverse this expanded space. Hence, these extensions are outside the scope of this study but are both potential research directions worth exploring.

The phylogenetic and hyperparameter estimates produced by SDPM analyses are averaged over the alignment partition space of rates and substitution pattern. These estimates therefore take into account the uncertainty associated with alignment partitioning. The user can therefore bypass the process of model and partition selection. Conversely, if one is interested in the across-site heterogeneity in the substitution process, our method can provide relevant information. Furthermore, it is clear from the large improvements in model fit that our approach goes some way to solving the problem of site to model assignment. We think that the methods described here provide a superior approach that can replace existing widely used methodologies for substitution model comparison and selection.

Supplementary Material

Supplementary figures S1–S10 and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the New Zealand Phylogenetics Meeting for fostering this work. They thank Dr. Simon J. Greenhill for his helpful suggestions. In addition, they thank Dr. David Posada and two anonymous reviewers for their very helpful comments on the manuscript. This work was supported by

Marsden Fund #UOA0809, a Rutherford Discovery Fellowship (to A.J.D.), a University of Auckland Doctoral Scholarship (to C.-H.W.) and NIH R01 GM086887 and R01 HG006139.

References

Antoniak CE. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann Stat.* 2:1152–1174.

Bruno WJ. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol Biol Evol.* 13:1368–1374.

Churchill GA, von Haeseler A, Navidi WC. 1992. Sample size for a phylogenetic inference. *Mol Biol Evol.* 9:753–769.

Dahl D. 2005. Sequentially allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models Technical report. Madison (WI): Department of Statistics, University of Wisconsin–Madison.

Dimmic M, Mindell D, Goldstein R. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac Symp Biocomput.* 5:18–29.

Drummond A, Nicholls G, Rodrigo A, Solomon W. 2002. Estimating mutation parameters, population history, and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161: 1307–1320.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22:1185–1192.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Ferguson TS. 1973. A Bayesian analysis of some nonparametric problems. *Ann Stat.* 1:209–230.

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* 97:611–631.

Fraley C, Raftery AE. 2006. Mclust version 3 for R: normal mixture modeling and model-based clustering. Technical Report 504. Seattle (WA): Department of Statistics, University of Washington.

Godsill SJ. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J Comput Graph Stat.* 10: 230–248.

Golding G. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol Biol Evol.* 1:125–42.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–458.

Gray R, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus O. 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol Biol.* 11:131.

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711–732.

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol.* 12:546–557.

Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst Biol.* 59:307–321.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22: 160–174.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Heled J, Drummond A. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27:570–580.

Höhna S, Defoin-Platel M, Drummond A. 2008. Clock-constrained tree proposal operators in Bayesian phylogenetic inference. Proceedings of the 8th IEEE International Conference on Bioinformatics and bioengineering, BIBE; 2008 October 8–10; Athens (Greece). p. 1–7.

Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. *Syst Biol.* 42:247–264.

Huelsenbeck JP, Jain S, Frost SWD, Pond SLK. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A.* 103:6263–6268.

Huelsenbeck JP, Joyce P, Lakner C, Ronquist F. 2008. Bayesian analysis of amino acid substitution models. *Philos Trans R Soc Lond B Biol Sci.* 363:3941–3953.

Huelsenbeck JP, Larget B, Alfaro ME. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol.* 21:1123–1133.

Huelsenbeck JP, Nielsen R. 1999. Variation in the pattern of nucleotide substitution across sites. *J Mol Evol.* 48:86–93.

Huelsenbeck JP, Suchard MA. 2007. A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol.* 56: 975–987.

Jeffreys H. 1998. Theory of probability. New York: Oxford University Press.

Jenkins G, Rambaut A, Pybus O, Holmes E. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol.* 54:156–165.

Jin L, Nei M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol Biol Evol.* 7:82–102.

Jukes T, Cantor C, Munro H. 1969. Mammalian protein metabolism. *Evol Protein Mol.* 3:21–132.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.

Koshi JM, Mindell DP, Goldstein RA. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. *Mol Biol Evol.* 16:173–179.

Kuo L, Mallick B. 1998. Variable selection for regression models. *Sankhya Indian J Stat Ser B (1960–2002).* 60:65–81.

Lanfear R, Calcott B, Ho S, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol.* 29:1695–1701.

Larget B, Simon D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol.* 16: 750–759.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.

Lemey P, Rambaut A, Drummond A, Suchard M. 2009. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* 5:e1000520.

Li P, Goldman N. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. *Mol Biol Evol.* 16: 1696–1710.

Mau B, Newton M. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *J Comput Graph Stat.* 122–131.

Mau B, Newton M, Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* 55:1–12.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys.* 21:1087.

Neal RM. 2000. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat.* 9:249–265.

Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J Royal Stat Soc Ser B.* 56:3–48.

Nielsen R. 1997. Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Syst Biol.* 46:346–353.

Olsen G. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp Quant Biol.* 52:825–837.

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol.* 53:571–581.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Pybus O, Charleston M, Gupta S, Rambaut A, Holmes E, Harvey P. 2001. The epidemic behavior of the hepatitis C virus. *Science* 292:2323.

Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A. 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol Biol Evol.* 20:381–387.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

R Development Core Team. 2011. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Rambaut A, Drummond AJ. 2009. Tracer. Available from: http://tree.bio.ed.ac.uk/software/tracer/ (last accessed December 6, 2012).

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M, Douzery E. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol.* 7:241.

Ray SC, Arthur RR, Carella A, Bukh J, Thomas DL. 2000. Genetic epidemiology of hepatitis C virus throughout Egypt. *J Infect Dis.* 182: 698–707.

Redelings B, Suchard M. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol.* 54:401–418.

Ronquist F, Teslenko M, van der Mark P, Ayres D, Darling A, Höhna S, Larget B, Liu L, Suchard M, Huelsenbeck J. 2012. Mrbayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.

Shapiro B, Rambaut A, Drummond A. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol.* 23:7–9.

Soyer O, Dimmic M, Goldstein R. 2002. Using evolutionary methods to study g-protein coupled receptors. *Pac Symp Biocomput.* 7: 625–636.

Suchard M, Kitchen C, Sinsheimer J, Weiss R. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst Biol.* 52:649–664.

Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol.* 18:1001–1013.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Ann Rev Ecol Evol Syst.* 36:445–466.

Swofford D, Olsen G, Waddell P, Hillis D. 1996. Phylogenetic inference. In: Hillis D, Moritz C, Mable B, editors. Molecular systematics, 2nd ed. Sunderland (MA): Sinauer Associates. p. 407–514.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 10:512–526.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci.* 17:57–86.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. 4th ed. New York: Springer.

Waddell P, Penny D. 1996. Evolutionary trees of apes and humans from DNA sequences. In: Lock A, Peters C, editors. Handbook of symbolic evolution. Oxford: Clarendon Press. p. 53–73.

Waddell PJ, Steel M. 1997. General time-reversible distances with unequal rates across sites: mixing and inverse Gaussian distributions with invariant sites. *Mol Phylogenet Evol.* 8:398–414.

Wertheim JO, Kosakovsky Pond SL. 2011. Purifying selection can obscure the ancient age of viral lineages. *Mol Biol Evol.* 28: 3355–3365.

Whelan S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol Biol Evol.* 25:1683–1694.

Wilson I, Balding D. 1998. Genealogical inference from microsatellite data. *Genetics* 150:499–510.

Wu C, Drummond AJ. 2011. Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov chain Monte Carlo. *Genetics* 188:151–164.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol.* 10: 1396–1401.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11:367–372.

Zlateva K, Lemey P, Moës E, Vandamme A, Van Ranst M. 2005. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. *J Virol.* 79: 9157–9167.